LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-668701

Mapping Conditions for Forced Migration

I. Abramova

March 18, 2015

Grace Hopper Celebration of Women in Computing
Houston, TX, United States
October 14, 2015 through October 16, 2015

Disclaimer

Mapping Conditions for Forced Migration

Irina Abramova
Lawrence Livermore National Laboratory*
abramova1@llnl.gov

ABSTRACT

In this work we use a large collection of international news articles to create a system for detection of early indicators of forced migration in a given geographic area. We integrate subject matter expert knowledge and data extracted from disparate sources of news articles. We leverage state-of-the-art Stanford NLP tools [1] to process the text corpus within the Apache Spark cluster-computing framework. We generate a knowledge concept graph and apply graph algorithms namely Random Walk with Restart (RWR) to measure the relatedness of concepts associated with forced migration and specific locations. The scoring of concepts across space and time aims to inform social scientists of troubling hotspots that indicate increased likelihood of forced migration.

AUDIENCE

[Data Science Applications], [Data Visualization], [Web Technologies], [Intermediate Technical Talk]

INTRODUCTION

The key questions we aim to address with this research are:

- Defining a data format to represent varied concepts such as people, geographic locations, and time as well as abstract concepts such as violence, displacement, government etc.
- Identifying concepts that are leading indicators of forced migration
- Geo-locating these concepts
- Defining model representation for the subject-matter expert knowledge and evaluating our results against these known concepts and events
- Quantifying the forced migration potential of a geographic location
- Developing a user interface for interactive visualization of migration potentials in geographic regions over time

We are using a dynamic concept graph as a core representation of concepts and their relationships over time. This representation provides the required flexibility and allows us to leverage existing distributed computation infrastructure based on Apache Spark/GraphX. For this project we picked a collection of 2.76M news articles related to Iraq and the Middle East, collected from the Georgetown EOS database [2] and extracted concepts as they relate to various regions and places in Iraq.

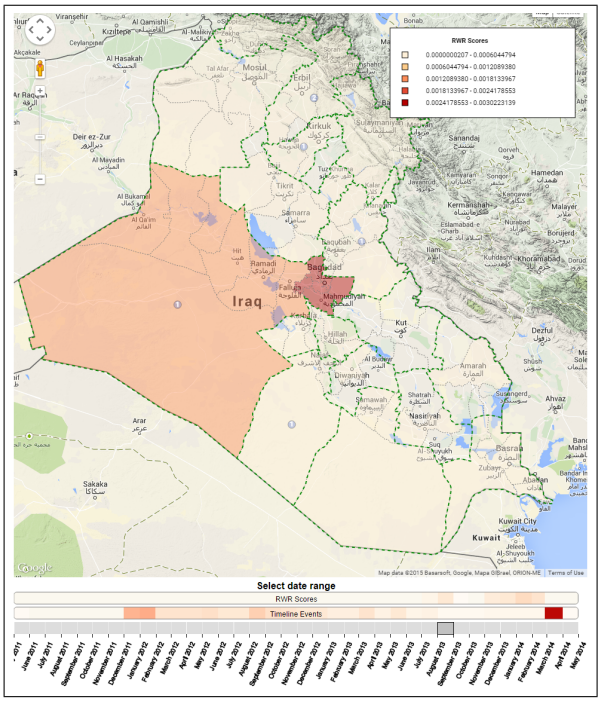

Fig. 1. Map widget shows Iraq regions highlighted according to the RWR score for violence

METHOD

The current pipeline to convert a collection of news articles to a dynamic concept graph consists of a few key steps.

We run Stanfords part-of-speech (POS) tagger to identify parts of speech and extract noun phrases (NP) using simple regex matching over the parts of speech tags. We ended up not using the Stanfords PCFG Parser for sentence structure parsing because we were seeing slow performance on larger sentences as well as some memory errors. We decided that regex parsing and matching of noun phrase structure is an acceptable alternative.

We create nodes in a graph out of the extracted noun phrases. The links are added between nodes if they happen

to be collocated within a window of 10 words. We capture extracted location entities as graph node attributes. The final graph consists of 27M vertices and 185M edges. To combat the problem of many noun phrase nodes containing non-unique locations we extract individual locations as special nodes in the graph to act as score sinks for multiple concepts.

We run a Random Walk with Restart (RWR) algorithm to associate concepts related to violence with geographic locations in Iraq. The resulting output of geographic locations and their scores with respect to violence is displayed on a map against a selection of space and time. Additionally, a list of relevant timeline events related to forced migration was compiled by the subject-matter experts. The list is then combined with the output of RWR algorithm.

TOPONYM RESOLUTION

The full dynamic graph generated contains a broad range of topics, concepts and entities that must be restricted to a given geographic area. Using the locations identified by the Stanford NE extractor we attempted to geo-locate the terms using GeoNames.org, a free geographical database containing over 10M structured records (elevation, population, latitude/longitude) for populated place names. However, querying this database for the country and region of a term, San Francisco for example, can produce many ambiguous results in multiple countries. Simple approaches, such as just using the first result, would be incorrect in the majority of cases. This prompted us to search for an alternative solution for the problem of geolocation.

We experimented with the CLAVIN Geocoder [3], a simple open-source tool for toponym resolution which does context resolution by looking at a sliding window of location data and resolves the terms based on the country code and region code commonality score. Toponym resolution can be significantly improved by incorporating recent advances in toponym resolution research. Using related data such as population statistics, provided the data is relevant and valid with respect to the given time interval, can help identify more prominent locations first. The concept of spatial minimality also reflects the general tendency of texts to only mention geographically near locations [4]. Using the spatial minimality principle the least ambiguous locations can be identified first, allowing for higher confidence in these term resolutions. The ambiguous terms can be identified on subsequent passes.

A simplification of the problem of toponym resolution can be done on limited data sets by leveraging pre-labeled data sources such as Wikipedia. For the current iteration of our work we are not attempting to perform location resolution. Instead we obtained a compiled list of region and city names in Iraq and performed matching and score calculation against the terms in this list.

RANDOM WALK WITH RESTART

In the original implementation, the list of location names was matched exactly against the node labels in the graph. This approach was easy to implement and simple to explain, but missed matches to nodes that included more just locations (e.g., protests in Baghdad), or multiple locations.

The original implementation also did not contain the temporal aspect, so the entire corpus was used to produce RWR scores regardless of the date the articles were published. In order to incorporate the time dynamics we split the graph by month intervals and performed RWR independently for each month.

We decided it made more sense to match on the location attributes of each node instead of the node label. Each node has a set of location attributes corresponding to the locations referenced in the label. This approach yielded far more matching nodes than the first approach (100 vs. 38K). However, this approach also has a major drawback, which is that many locations occur in multiple nodes. This means that we end up with many RWR scores for a popular location such as Baghdad. There is not an obvious way to aggregate these multiple scores in a principled way to get a single score for each location.

Based on the above, we tried a third approach, which involves a small change to the ontology and structure of the concept graph. For every location that occurs as a location attribute of an NP node, we create a new LOCATION node and a directed edge from the NP node to the LOCATION node. This representation has the advantage that there is exactly one LOCATION node in the graph for each location name. So, there is an obvious 1-to-1 mapping between RWR scores and location names. Note that in this representation LOCATION nodes only have incoming links. This is to prevent LOCATIONs from connecting otherwise disconnected NPs (i.e., NPs that do not co-occur); and LOCATIONs from connecting NPs across time, causing problems for the temporal RWR version. This approach is marginally more complicated, but the basic idea is still straightforward: we have the main NP co-location graph on which random walks occur only within a specified time period and we have terminal LOCATION nodes that end up with a portion of the mass that flows from the NP nodes as the random walks occur.

USER INTERFACE

We implemented a dashboard web application along with a collection of widgets that allow for navigating the scores output from the RWR algorithm. The dashboard provides a widget to perform a faceted search over the original collection of documents. We also developed a mapping widget that contains a geographic area of interest and displays a list of scored concepts as a heatmap of highlighted regions.

The users are presented with a geographic map of Iraq and a time slider for a specific range of dates. The list of scores and geographic locations are geocoded and mapped to a corresponding province/region of Iraq. Each of the regions is highlighted according to a score it received with respect to the concept of violence for the currently selected time interval. The highlighted regions color intensity is increased as the violence score increases.

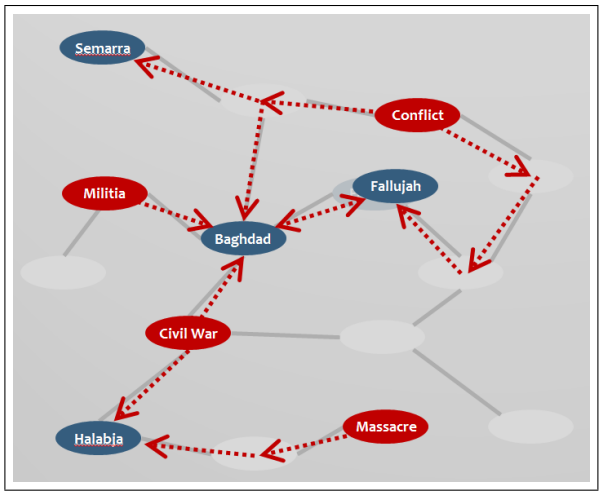

Fig. 2. Using Random Walk with Restart to associate violence with locations

The document search widget allows for browsing the original documents used to generate the concept graph. Alternatively, it can be used to view the documents that were used to produce the RWR score for the specified region and time interval.

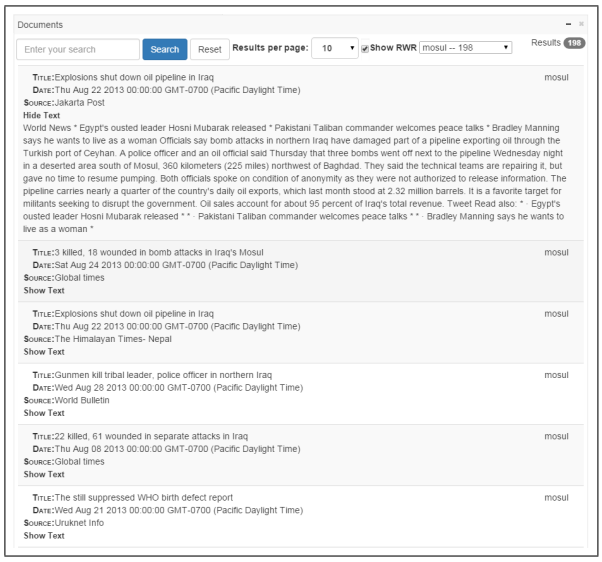

Fig. 3. Document search widget allows the user to browse original news articles and navigate from RWR scores to the text passages that generated them

OUTCOMES/CONCLUSIONS

The goal of this work is to discover areas and times for which local conditions like violence, break-down in the rule of law, hunger, and disease increase the local probability of migration. By translating a corpus of documents into a dynamic concept graph we can identify relevant concepts and perform query expansion with concept geolocation.

We are still investigating the classes of nodes that can most effectively represent structured knowledge relevant for forced migration potential. We are refining our concept node identification by applying techniques like tf.idf, bunching statistics, and running pagerank [5]. This will allow us to identify key phrases in a document and perform some pre-filtering to capture the most informative nodes.

The concept of an edge is another open area of research. Besides word co-location, an edge can also mean syntactic relationships between words or geometric relationships in semantic space. By capturing other types of edges or employing the concepts of semantic spaces [6] we can merge similar nodes and adjust meaning and representation of the concept graph.

This project is an ongoing research effort with funding for the next year.

PARTICIPATION STATEMENT

I will attend the conference if accepted.

BIO

Irina Abramova is a computer scientist at Lawrence Livermore National Laboratory working in the Global Security Directorate. She earned her Bachelors of Science in Computer Science from University of Southern California and an Artificial Intelligence Graduate Certificate from Stanford University. Irina's areas of expertise are web application design and development, data visualization, user interface design and unstructured text processing. She enjoys working with the Grails web framework, JavaScript technologies such as D3.js, Backbone MVC, jQuery etc. Irina enjoys trying new technologies and took the opportunity to explore Spark and Scala while working on this project.

ACKNOWLEDGMENT

I want to mention James Brase, David Buttler, Brian Gallagher, Kyle Halliday and Eric Wang as members of this project who have made significant contributions to this work and provided assistance in writing this paper.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

REFERENCES

[1] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014.The Stanford CoreNLP Natural Language Processing Toolkit. InProceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[2] Forecasting the Break: Building Community and Capacity for Large-Scale Data-Intensive Research in Forced Migration Studies. (n.d.). Retrieved from http://isim.georgetown.edu/forecasting

[3] "Berico-Technologies/CLAVIN." Github. Bericho-Technologies, 21 Oct. 2014. Retrieved from https://github.com/Berico-Technologies/CLAVIN

[4] Speriosu, Michael Adrian. "Methods and applications of text-driven toponym resolution with indirect supervision." 2013. Ph.D. thesis. University of Texas

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of NIPS, 2013.

[6] Mihalcea, R., Tarau, P.. TextRank: Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004